SUPPLEMENTARY FIGURE LEGENDS

Figure S1. No. of fusions detected in different tissue/cell types. The majority of fusions were unique to one particular tissue/cell type. A little over 10% of fusions were seen in more than one sample type. *CTBS-GNG5* was found in 15 different tissues and cell lines. *TIMM23B-LINC00843* was observed in 26 different tissues and cell lines.

Figure S2. Chimeric peptide identification flowchart. About 200bp sequences across the fusion junctions were translated into three reading frames. Peptides with less than 20 amino acids on either side were filtered off. *In silico* tryptic digestions yielded peptides with R or K on the ends. These peptide sequences plus Uniprot. Database were used as input. We then performed tblastn to further filter off peptides that can also be products of reference genes.

Figure S3. The density of the fusion-forming genes per mega base correlates with the overall density of genes on individual chromosomes.

Figure S4. Fusion classification according to the junction position relative to the exon of parental genes. E-known exon boundary. M-middle of known exon. All the possible outcome for fusions are E/E, E/M, M/E or M/M. Red blocks represent exons of the 5' gene. Green blocks represent exons of the 3' gene.

Figure S5. Distribution of M/M fusions according to the length of Short Homologous Sequences (SHS).

Figure S6. Known protein-binding motifs matching the motifs found through MEME. Using p=0.001 as cutoff, five known motifs were found similar to the 5' gene downstream motif. One known motif was found similar to the 3' gene upstream motif.

SUPPLEMENTARY TABLES

Table S1. Summary of RNA–sequencing libraries.

Table S2. List of all 11,531 candidate fusion RNAs. Tissue sources, parental gene positions, fusion locus, fusion type, and protein-coding potentials are included.

Table S3. List of recurrent fusions. Parental gene location and direction on individual chromosomes are listed according to hg19. Highlighted in yellow are the ones that confirmed by RT-PCR and Sanger sequencing. Red fonted are the fusions confirmed by Western blot or Mass Spec.

Table S4. 51 fusions found in more than five tissue/cell types.

Table S5. Chimeric peptides supported by LCMS of MCF10A cells. Included are peptide sequence and corresponding fusions.

Table S6. Gene ontology terms for parental genes identified by GOrilla. Sheet1 is for recurrent fusions. Sheet2 is for tissue-specific fusions. Highlighted in yellow are the GO terms unique to recurrent fusion parental genes.

Table S7. Recurrent mouse fusion RNAs. Highlighted in yellow are the gene pairs that are conserved in human.

Table S8. Common fusions in the normal tissues/cells and Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.

Table S9. Gene Ontology terms enriched in "non-cancer", "cancer", and "both" fusions. Sheet 1 is for parental genes involved in "non-cancer" fusions only. Sheet2 is for parental genes involved only in the "cancer" fusions. Sheet3 is for parental genes involved in both categories.